Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 57-64 recibido 04-04-2018 revisado 02-05-2018 aceptado 04-05-2018

Improving the accessibility of biomedical texts by semantic enrichment and definition expansion

Mejora de la accesibilidad de textos biomédicos mediante enriquecimiento semántico y expansión de definiciones

Pablo Accuosto and Horacio Saggion
Large-Scale Text Understanding Systems Lab
TALN Research Group, DTIC
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
{*name.surname*}@upf.edu

Abstract: We present work aimed at facilitating the comprehensibility of health-related English-Spanish parallel texts by means of the semantic annotation of biomedical concepts and the automatic expansion of their definitions. In order to overcome the limitations posed by the scarcity of resources available for Spanish, we propose to exploit existing tools targeted at English and then transfer the produced annotations. The evaluations performed show the feasibility of this approach. An enriched set of texts is made available, which can be retrieved, visualized and downloaded through a web interface.
Keywords: semantic annotation, definition expansion, biomedical terminology

Resumen: Este trabajo busca facilitar la comprensión de textos médicos en un corpus paralelo inglés-español mediante la anotación semántica de conceptos y la expansión automática de definiciones. Considerando la limitación de recursos disponibles para el español, proponemos explotar herramientas dirigidas al inglés para obtener anotaciones que luego se transfieren a los textos en español. Las evaluaciones realizadas muestran la viabilidad de este enfoque. Se hace público un conjunto de textos enriquecidos que se pueden recuperar, visualizar y descargar mediante una interfaz web.
Palabras clave: anotación semántica, expansión de definiciones, terminología biomédica

1 Introduction

A vast volume of biomedical knowledge is generated on a daily basis as unstructured text, including scientific articles (Bornmann and Mutz, 2015), patents and medical reports. Natural language processing (NLP) tools have a great deal to contribute to the effective exploitation of this knowledge. In particular, the automatic annotation of biomedical texts with concepts from manually curated thesaurus and knowledge bases, such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), is key to make biomedical knowledge manageable and discoverable. At the same time, linking complex terms in health-related texts to uniquely identified concepts makes it possible to expand their definitions and/or enrich them with information which can improve the text comprehensibility for general audiences. This is particularly relevant as studies show that the possibility of understanding health-related information ("health literacy") predicts a person's health status more accurately than variables such as age, income, level of education and race (MacLeod et al., 2017).

Several resources, tools and methods have been developed to support and/or automate the semantic indexing of biomedical texts, as recently reviewed by (Jovanović and Bagheri, 2017). Most of these resources are, nevertheless, only available for English, with a few exceptions described in the cited review. The lack of tools with similar levels of maturity for Spanish exacerbates an existing imbalance in the access to health-related information for speakers of this language.

ISSN 1135-5948. DOI 10.26342/2018-61-6

The work described in this paper seeks to enhance the access to health information by non-expert population by means of semantic indexing and enrichment of biomedical texts in Spanish. Our hypothesis is that adding a layer of semantic information to biomedical texts can contribute to make them more accessible, as automatically retrieved definitions and related information can be made available to facilitate the comprehensibility of technical terms by non-expert users.

1.1 Contributions

Our contributions can be summarized as:

- We explore the possibility of exploiting existing resources and off-the-shelf tools available for English for the annotation of biomedical texts in Spanish;
- We propose a method for transferring automatically-obtained annotations in English texts to their parallel Spanish versions, which we evaluate against a gold standard biomedical corpus. To the extent of our knowledge, our work is the first to compare, for this task, the performance of two semantic similarity functions: one that relies on traditional information retrieval methods based on TF-IDF sparse vectors and another one based on dense vectors that include subword information.
- In order to assess the viability of our approach, we develop a prototype for the annotation of the ScieLO parallel corpus (Neves, Jimeno-Yepes, and Névéol, 2016) as well as an on-line tool that allows the search and visualization of semantically enriched documents in English and Spanish;
- The linguistic resources generated in the context of this work, including the annotated documents in JSON format, are made available for download from the AsisTerm web site.[1]

2 Related work

The automatic or semi-automatic annotation of biomedical texts in English has gathered considerable attention in the past decade and several tools and resources have been developed in this area including cTAKES,[2] MetaMap,[3] NCBO Annotator,[4] and BeCAS,[5] among others. We refer the reader to (Jovanović and Bagheri, 2017), (Hassanzadeh, Nguyen, and Koopman, 2016) and (Groza, Oellrich, and Collier, 2013) for detailed descriptions of these systems and their performance.

The systems developed in the context of the 2013 CLEF-ER challenge for biomedical entity recognition in parallel multilingual corpora (Rebholz-Schuhmann et al., 2013) provide some of the first prototype tools for the annotation of biomedical texts in languages other than English. Among the participating systems there were some targeted at Spanish including the proposed by (Bodnari et al., 2013) and (Attardi, Buzzelli, and Sartiano, 2013), which exploited word alignment information obtained by statistical translation tools in order to transfer annotations from English to Spanish that were then used to train named-entity taggers. In turn, (Berlanga, Nebot, and Jimenez, 2010) introduced the notion of *concept retrieval*, which was based on applying information retrieval methods in order to obtain UMLS concepts relevant to a text and later use them to properly annotate matching text spans. Other initiatives aimed at automatically annotating Spanish biomedical texts include (Carrero, Cortizo, and Gómez, 2008), who proposed to combine machine translation and the MetaMap in order to annotate Spanish texts with UMLS concepts, and (Castro et al., 2010), who developed an automatic system for the recognition of SNOMED CT concepts by computing a similarity function between sentences in clinical notes and SNOMED CT concepts based on the results obtained by querying an Apache Lucene[6] index. (Oronoz et al., 2013) extended the Freeling Spanish analyzer[7] to recognize biomedical entities extracted from available knowledge resources (lists of medical abbreviations and drug names, as well as the SNOMED CT thesaurus) and, more recently, (Pérez, Cuadros, and Rigau, 2018) developed a prototype that uses the UMLS Metathesaurus for biomedical term normalization in order to enrich electronic health records in Spanish.

[1] http://scientmin.taln.upf.edu/scielo/
[2] http://ctakes.apache.org/
[3] https://metamap.nlm.nih.gov/
[4] http://bioontology.org/annotator-service
[5] http://bioinformatics.ua.pt/becas
[6] https://lucene.apache.org/
[7] http://nlp.lsi.upc.edu/freeling/

They evaluated their system by measuring the agreement obtained in parallel English-Spanish corpora annotated with MetaMap (for English) and their prototype (for Spanish). In order to do this, they annotated a set of Spanish health records and their English translations, as well as a manually revised version of the ScieLO corpus. Due to the differences between the proposed approaches and the evaluation methods and datasets–in the cases in which evaluation results are made available–, none of the results of these previous works can be directly compared to ours.

3 Semantic annotation of an English-Spanish parallel corpus

3.1 The ScieLO parallel corpus

The ScieLO English-Spanish parallel corpus contains 17,015 metadata entries in Dublin Core XML format from documents included in the SciELO collection, a database of open-access scientific publications. The corpus covers a collection of Spanish health-related scientific journals selected based on their quality. The metadata includes publication information in Spanish (venue, keywords, authors, date) and bi-lingual (English and Spanish) versions of the titles and the abstracts.

3.2 Corpus indexing

In order to implement efficient full search and retrieval functionalities in English and Spanish the ScieLO abstracts were converted from XML to JSON and indexed with the Elasticsearch search engine.[8] Basic language processing of the texts (stemming, stop-word removal, relevance scoring of terms) was performed at indexing by means of the standard English and Spanish text analyzers included in Elasticsearch.[9]

3.3 Annotation of English abstracts

The semantic annotation of the ScieLO abstracts in English was done by means of the BeCAS annotation services' API.[10] The choice of BeCAS as off-the-shelf annotation system responded mainly to its ease of use and acceptable performance when evaluated against the Medline titles included in the English-Spanish Mantra gold standard (see Section 4.1).

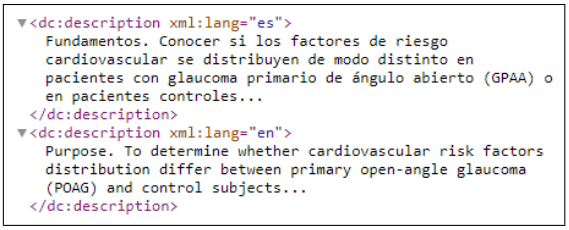

Figure 1: Fragment of a ScieLO document

In the current prototype a filter is applied when calling the BeCAS service in order to retrieve annotations corresponding to the semantic group "DISO" (*Disorders*). We foresee, in future experiments, to expand the coverage of the annotations to the groups *Anatomy*, and *Biological processes*.

We will also analyze the results obtained with other annotation tools with a broader coverage of UMLS types. In particular, the NIH MetaMap, which covers 134 semantic types, in contrast to the 26 UMLS types included in BeCAS.

Even if not an essential element in our proof-of-concept system, the choice of semantic annotator for English would clearly be determining in a real-world case scenario, as the quality and coverage of the annotations of the English text sets an upper bound for the overall performance of the system.

3.4 Transfer of annotations to abstracts in Spanish

Once the relevant UMLS concepts are identified in an English abstract by means of the BeCAS service, the spans of texts in the parallel Spanish abstract that best match each of them have to be determined in order to transfer the annotations. Consider, for example, the following fragment from the example included in Fig. 1.

... risk factors distribution differ between ***[primary open-angle glaucoma]*** *(****[POAG]****) and control subjects. To assess the strength of this association in* ***[POAG]****.*

BeCAS, in this case, correctly associates the UMLS concept *C0339573* to the three spans of texts in bold. We would like each of these instances to be associated to the corresponding text spans in the Spanish version:

... los factores de riesgo cardiovascular se distribuyen de modo distinto en pacientes con ***[glaucoma primario de ángulo abierto]*** *(****[GPAA]****) o*

[8] https://www.elastic.co/

[9] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-analyzers.html

[10] http://bioinformatics.ua.pt/becas/api

*en pacientes controles. Cuantificar la prevalencia de estos factores en el **[GPAA]**.*

We assume that the instances of the same concepts appear in the same order in the English and Spanish texts. Therefore, we process the Spanish abstract sequentially to find, for each identified concept instance, the text span in Spanish that best matches it. In order to do this, we compute the similarity between each considered text span and all the lexicalizations of the concept available in UMLS.[11] Once a concept instance is associated to a text span, we consider its final offset to continue looking for the next one.[12] It might be the case that, for a given instance of a concept, no matching text span can be identified. In this case, the system will continue with the next annotation retrieved by BeCAS. In the example this might happen, for instance, if the system could not associate the first occurrence of the acronym "GPAA" to the corresponding concept (*primary open-angle glaucoma*). In this case, we would like the system to drop that particular instance of the concept and continue processing the text, looking for a span to which associate the following instance (in this case, the second occurrence of the term "GPAA"). We therefore define a search window for term candidates in Spanish based on the relative position of the annotation in the English text. Defining a search window allows us to keep some level of alignment between the two texts, which minimizes the possibility of associating annotations to the wrong occurrence of a term while, at the same time, prevents the system to keep consuming concept instances in a given position without moving forward.

3.4.1 Candidate terms generation

In order to identify spans of text in Spanish as candidates for being annotated, we first split the abstract into sentences, tokenize it and perform part-of-speech (POS) tagging[13] by means of the Stanford CoreNLP library[14] (Manning et al., 2014).

In order to cope with errors produced by the POS tagger and grammatical errors in the source texts, we do not restrict the potential text spans to well-formed noun phrases but instead allow some flexibility in the sequences of POS that can be considered to constitute candidate terms. In fact, we allow any sequence of up to eight tokens beginning with a token tagged as NOUN or PROPN, ending with a token tagged as NOUN, PROPN, ADJ, VERB or NUM, and containing tokens tagged with NOUN, PROPN, ADJ, DASH,[15] CONJ, ADP, DET, ADV, VERB or NUM as potential term candidates. These heuristics are based on the most frequent POS sequences occurring in UMLS terms. In the example above, this rule would produce, as candidate terms: *glaucoma*, *glaucoma primario*, *glaucoma primario de ángulo*, *glaucoma primario de ángulo abierto*.

3.4.2 Similarity computation

As mentioned in Section 3.4, a similarity score is computed between candidate terms generated at a given offset in the Spanish text and the UMLS concepts retrieved by BeCAS from the English version. We evaluated two similarity functions: the first one is Elasticsearch's implementation of the BM25 ranking function, which computes the cosine similarity of TF-IDF vectors obtained from the normalized terms of a query (in our case, the candidate text span) and a document (in our case, all the Spanish variants of the concept in UMLS).[16] The second scoring function considered is the cosine similarity between dense vector representations of candidate terms and the Spanish lexicalizations of UMLS concepts. Both for the candidate terms and the lexicalizations of UMLS concepts, their corresponding dense vectors are computed as the average of the normalized embeddings of the words included in them. For the word embeddings we used fastText vectors (Bojanowski et al., 2016) pre-trained with Wikipedia pages in Spanish.[17] Since word representations in fastText are computed as the sum of their character n-gram vectors, embeddings for out-of-vocabulary words can be generated on the

[11] As shown in Table 1, among the sources currently being considered, the one with more concepts variants in Spanish is SNOMED CT, with over 1 million entries.

[12] Overlapping and nested annotations are not considered in the current prototype.

[13] Using a slightly modified version of the universal POS tags: `http://universaldependencies.org/`

[14] `https://stanfordnlp.github.io/CoreNLP/`

[15] The DASH tag was added as the character "-" is frequently used as part of biomedical terms.

[16] `https://www.elastic.co/guide/en/elasticsearch/guide/current/practical-scoring-function.html`

[17] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

Source	UMLS code	Entries
CPT	CPTSP	2,707
ICPC	ICPCSPA	723
LOINC Argentina	LNC-ES-AR	76,586
LOINC Switzerland	LNC-ES-CH	4,940
LOINC Spain	LNC-ES-ES	52,641
MedDRA	MDRSPA	102,097
MeSH	MSHSPA	70,033
SNOMED CT	SCTSPA	1,084,815
WHO ART	WHOSPA	3,106

Table 1: Metathesaurus sources in Spanish with number of concept variants

fly. Considering subword information allows the embeddings-based function to correctly assign a high similarity score (0.93) when comparing the concept *C0025179* (lexicalized in the Spanish UMLS as *Metilglucamina* and *Meglumina*) and the candidate term *N-metil glucamina*, which occurs in the Mantra Medline corpus used to evaluate our proposal (see Section 4). This similarity is not captured by the Elasticsearch-based function, which assigns a score of 0 to the candidate term.

4 Evaluation

The Mantra project[18] was one of the first initiatives aimed at the multilingual processing of biomedical texts. In its context, valuable resources were generated, including the Mantra gold-standard parallel corpora for biomedical concept recognition (Mantra GSC) (Kors et al., 2015). The Mantra GSC consists of three parallel corpora: Medline titles, sentences from drug labels provided by the European Medicines Agency (EMEA), and sentences from patents made available by the European Patent Office. The Medline and EMEA corpora include parallel texts in English, French, Dutch, German and Spanish, while the patents corpus is available for English, French and German. In the case of the English-Spanish pairs, both Medline and EMEA corpora include 100 textual parallel units (titles or sentences) annotated with a subset of UMLS concepts from MeSH,[19] SNOMED CT,[20] and MedDRA.[21]

We evaluated the feasibility of the proposed approach for automatically transferring English annotations to Spanish texts by comparing the annotations produced by our system with those included in the Mantra GSC.

For all the evaluations we computed precision (P), recall (R) and F1-score for exact text spans (same boundaries in the gold standard and automatic annotations) as well as for overlapping spans. In order to assess the loss of accuracy when non-exact matching spans where considered, an "overlapping percentage" (OP) was calculated as the relation between the length of the overlapping span and the length of the longest span between the annotation produced by our system and the one in the gold standard. We also considered in our evaluations whether the same gold standard concept CUIs were identified or if different ones where produced by the automatic annotation system. The less restrictive alternatives (overlapping spans and non-matching concepts) were evaluated since they can be of use in specific applications, as argued by (Hassanzadeh, Nguyen, and Koopman, 2016). We report below the results obtained considering only when gold standard CUIs are identified, for the sake of space.[22]

BeCAS does not produce discontinuous annotations and only continuous candidate text spans were considered in the texts in Spanish. Therefore, only continuous annotations in the Mantra corpora were considered for the evaluation and including the first continuous portion of discontinuous gold standard annotations.[23]

The parallel Medline and EMEA corpora included in Mantra are each annotated with 64 different semantic types but BeCAS currently includes only 26.[24] In order to make both sets comparable we considered for the evaluation only Mantra annotations with semantic types recognized by BeCAS.

4.1 Evaluation of the annotations produced by BeCAS

We were interested in assessing both the transferring process and the outcomes of the full processing pipeline, including the automatic annotation of the English texts by means of the BeCAS service. We evaluated independently the annotations produced by

[18] `https://sites.google.com/site/mantraeu/`

[19] `https://www.nlm.nih.gov/mesh/`

[20] `https://www.snomed.org/snomed-ct`

[21] `https://www.meddra.org/`

[22] The full results are available at `http://scientmin.taln.upf.edu/scielo/evaluations/Evaluations_AsisTerm.pdf`

[23] Note that 98% of the annotations in the Medline and EMEA corpora are continuous.

[24] `http://bioinformatics.ua.pt/becas/about`

BeCAS against the Mantra English corpora, as it establishes an upper bound for the results of the full pipeline. The F1-score obtained for the BeCAS annotations when considering matching span boundaries was of 0.76 in the case of the Medline corpus and of of 0.67 in the case of EMEA.

4.2 Evaluation of the annotation transferring process

Table 2 shows the P, R and F1 results obtained when evaluating the transfer of annotations between the English and the Spanish versions of the Mantra GSC. The F1-scores obtained for the exact matching boundaries for Medline were of 0.60 for the embeddings-based similarity function (FT) and of 0.57 for the Elasticsearch-based one (ES). Different thresholds were considered to decide whether there was a valid match between a candidate term and a concept string when comparing the fastText embeddings. In the case of Medline the best results were obtained with a threshold of 0.7875. In the case of EMEA, the F1-scores obtained were of 0.60 for the embeddings-based similarity function (with a threshold of 0.9250) and of 0.59 for Elasticsearch. It is relevant to note that, even if the embeddings are expected to better capture semantic similarities between candidate text spans and UMLS concepts, considering all the lexicalizations available from multiple UMLS sources for a given concept contribute to mitigate this advantage, which yields to obtaining competitive results with sparse vectors. The choice of similarity function heavily depends, therefore, on the intended usage of the system and implementation decisions.

Due to length restrictions, it is not possible to include in this paper a detailed error analysis. The assumption that the same set of annotations appear in the same order in the English and Spanish versions of the texts and processing them sequentially in the order in which they appear in English, as implemented in our simplified prototype, can explain some of the errors. Others are originated by the proposed similarity functions failing to identify the right candidate term for a given concept. In general, in the case of the Elasticsearch function, this is due to lexical differences between the terms appearing in the texts and the variants of the concepts included in UMLS, as shown in the example

Similarity	**Span**	**P**	**R**	**F1**	**OP**
Medline corpus					
FT	exact	0,77	0,49	0,60	
FT	overlap	0,93	0,59	0,72	0,93
ES	exact	0,74	0,47	0,57	
ES	overlap	0,95	0,62	0,75	0,91
EMEA corpus					
FT	exact	0,89	0,45	0,60	
FT	overlap	0,95	0,49	0,65	0,98
ES	exact	0,77	0,48	0,59	
ES	overlap	0,91	0,58	0,71	0,92

Table 2: Evaluation of annotations transferred from English to Spanish Mantra GSC

Similarity	**Span**	**P**	**R**	**F1**	**OP**
Medline corpus					
FT	exact	0,69	0,58	0,63	
FT	overlap	0,75	0,67	0,71	0,94
ES	exact	0,64	0,51	0,57	
ES	overlap	0,74	0,64	0,69	0,91
EMEA corpus					
FT	exact	0,67	0,49	0,57	
FT	overlap	0,61	0,52	0,56	0,99
ES	exact	0,56	0,49	0,52	
ES	overlap	0,64	0,65	0,65	0,89

Table 3: Evaluation of the full pipeline applied to the Mantra GSC

mentioned in section 3.4.2. In the case of the embeddings-based function some errors can be explained by the difficulty to establish a unique, good-for-all, similarity threshold that provides a correct balance between precision and recall so as to identify, in one pass, the best term candidates and their exact boundaries. Performing multiple iterations over the candidate terms–possibly considering different similarity thresholds in each iteration in the case of the embeddings-based score–could contribute to partially mitigate these errors.

4.3 Evaluation of the full pipeline

Table 3 shows the results of evaluating the annotations obtained when applying the full processing pipeline, including the automatic annotation of the English texts produced by BeCAS and their transfer to the Spanish versions. When transferring the annotations obtained with BeCAS for the Medline corpus we obtained F1-scores of 0.63 and 0.57 for the fastText and Elasticsearch similarity functions, respectively (consider-

ing exact span boundaries), whereas for the EMEA corpus the F1-scores were of 0.57 and 0.52 for fastText and Elasticsearch functions, respectively. As there were not significant differences between the Medline and EMEA corpora in the evaluation of the annotation transferring process, these lower F1-scores could be explained by the poorer performance observed when annotating EMEA with BeCAS, as mentioned in Section 4.1.

5 The AsisTerm prototype

AsisTerm[25] provides an on-line interface to search and visualize biomedical abstracts from the ScieLO parallel corpus (Neves, Jimeno-Yepes, and Névéol, 2016) in English and Spanish annotated with UMLS concepts.

The annotated abstracts can be downloaded as JSON files including, for each annotation, its starting and ending offsets, the annotated text, and the corresponding UMLS concept, including its CUI, semantic type and group in the Metathesaurus.[26]

5.1 Definition expansion

One of the main objectives for associating annotations to the ScieLO abstracts was to make it possible to retrieve additional information that can contribute to facilitate the comprehension of complex terms included in them. When a Spanish abstract is displayed on AsisTerm, its annotations are retrieved and the corresponding text spans are highlighted. When the user clicks on an highlighted text span, all the source-specific identifiers and lexicalizations associated to the concept are displayed, as well as their corresponding definitions, if available. In most cases, UMLS concepts do not have definitions associated in the Metathesaurus.[27] In order to overcome this limitation, we retrieve the definitions (and/or additional information related to the concept) by means of the MedlinePlus Connect API,[28] querying it by the concept's SNOMED CT code.[29] When available (currently for SNOMED, MeSH and MedlinePlus), additional links are included to external pages with source-specific information, such as the concepts' hyperonyms, synonyms, and hyponyms (in the case of SNOMED or MeSH) or related information (in the case of MedlinePlus).

6 Conclusions and future work

In this paper we have presented a prototype aimed at semantically indexing complex terms in biomedical texts as a first step to improve their comprehensibility. As a proof-of-concept experiment, we used these annotations to retrieve and display definitions and related information. We applied our proposal to the annotation of the ScieLO English-Spanish parallel corpus and developed a web-based system to allow searching and visualizing its enriched contents in English and Spanish. We presented a proposal for exploiting existing tools targeted at English and transferring the obtained annotations to Spanish, comparing the performance obtained by means of a classic information retrieval similarity ranking function and the cosine similarity in a continuous vector space. We evaluated both approaches with English-Spanish gold standard corpora in the biomedical domain, obtaining promising results.

In terms of potential extensions to our work, we would like to investigate whether a harmonized combination of annotations obtained from multiple existing tools would significantly improve the accuracy of the results. We would also like to analyze the results obtained with embeddings trained with biomedical texts, which should contribute to obtain vectors better suited for this particular task. Another area to explore is the combination of our annotation transferring proposal with machine translations tools, which would allow to use the system in contexts where no parallel texts are available. Finally, we are also particularly interested in conducting usability tests to assess the degree to which the enriched texts can effectively contribute to improve the comprehension of complex texts by non-expert users.

Acknowledgements

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the TUNER project (TIN2015-65308-C5-5-R, MINECO / FEDER, UE).

[25] `http://scientmin.taln.upf.edu/scielo/`

[26] `http://ncbi.nlm.nih.gov/books/NBK9679/`

[27] For the subset of UMLS sources that we are currently working with, only 32,338 concepts have definitions in English and 7,154 have definitions in Spanish.

[28] `https://medlineplus.gov/connect/`

[29] MedlinePlus Connect supports queries by SNOMED CT and ICD-10-CM codes.

References

Attardi, G., A. Buzzelli, and D. Sartiano. 2013. Machine translation for entity recognition across languages in biomedical documents. In *CLEF (Working Notes)*.

Berlanga, R., V. Nebot, and E. Jimenez. 2010. Semantic annotation of biomedical texts through concept retrieval. *Procesamiento del Lenguaje Natural*, 45:247–250.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bodnari, A., A. Névéol, Ö. Uzuner, P. Zweigenbaum, and P. Szolovits. 2013. Multilingual named-entity recognition from parallel corpora. In *CLEF (Working Notes)*.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bornmann, L. and R. Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Carrero, F., J. C. Cortizo, and J. M. Gómez. 2008. Building a Spanish MMTx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 346–353. Springer.

Castro, E., A. Iglesias, P. Martínez, and L. Castano. 2010. Automatic identification of biomedical concepts in Spanish-language unstructured clinical texts. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 751–757. ACM.

Groza, T., A. Oellrich, and N. Collier. 2013. Using silver and semi-gold standard corpora to compare open named entity recognisers. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 481–485.

Hassanzadeh, H., A. Nguyen, and B. Koopman. 2016. Evaluation of medical concept annotation systems on clinical records. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 15–24.

Jovanović, J. and E. Bagheri. 2017. Semantic annotation in biomedicine: the current landscape. *Journal of biomedical semantics*, 8(1):44.

Kors, J. A., S. Clematide, S. A. Akhondi, E. M. Van Mulligen, and D. Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.

MacLeod, S., S. Musich, S. Gulyas, Y. Cheng, R. Tkatch, D. Cempellin, G. R. Bhattarai, K. Hawkins, and C. S. Yeh. 2017. The impact of inadequate health literacy on patient satisfaction, healthcare utilization, and expenditures among older adults. *Geriatric Nursing*, 38(4):334–341.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Neves, M. L., A. Jimeno-Yepes, and A. Névéol. 2016. The ScieLO corpus: a parallel corpus of scientific publications for biomedicine. In *LREC*.

Oronoz, M., A. Casillas, K. Gojenola, and A. Perez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*, pages 536–543. Springer.

Pérez, N., M. Cuadros, and G. Rigau. 2018. Biomedical term normalization of EHRs with UMLS. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Rebholz-Schuhmann, D., S. Clematide, F. Rinaldi, S. Kafkas, E. M. van Mulligen, C. Bui, J. Hellrich, I. Lewin, D. Milward, M. Poprat, et al. 2013. Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 353–367. Springer.